Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 20, No. 3, 2026

A Review of Colorectal Polyp Segmentation Methods

Wenwu Zhang, Baishun Su, Caihong Huangfu, Lihong Zhang*

College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454003, China
*** Corresponding author**

Abstract: Colorectal cancer is a malignant tumor with a persistently high incidence worldwide, and the majority of cases arise from the malignant transformation of polyps. However, polyps in colonoscopic images exhibit significant variations in morphology and scale, and the boundaries between polyps and surrounding tissues are often indistinct, which makes manual detection particularly challenging. On the one hand, the diagnostic process relies heavily on physicians' experience, which may lead to missed diagnoses and misdiagnoses. On the other hand, the limited number of experienced clinicians makes it difficult to meet patients' needs in a timely manner. This diagnostic uncertainty, combined with the scarcity of medical resources, highlights the critical value of AI-assisted diagnostic systems. Developing high-performance polyp segmentation algorithms that accurately locate and delineate polyp regions can improve diagnostic efficiency and provide effective decision support for clinicians. Recent advances in deep learning have significantly improved colorectal polyp segmentation. Regarding network architectures, models have evolved from early U-shaped designs to Transformer-based architectures that capture global context. More recently, Mamba-like networks and dual-branch architectures have also been proposed. In terms of research paradigms, this paper focuses on the application of methods such as multi-scale feature fusion, attention mechanisms, and diffusion models. In addition, this study summarizes commonly used datasets in the field of polyp segmentation and provides a detailed description of evaluation metrics closely related to segmentation performance. Finally, this paper systematically analyzes the limitations of current colorectal polyp segmentation methods in clinical practice. Based on this analysis, it discusses potential directions and trends for future algorithm development, providing a reference for further research.

Keywords: Colorectal polyp; Deep learning; Segmentation method; Learning paradigm; Dataset.

1. Introduction

Colorectal cancer, a type of malignant tumor, is the third most common cancer worldwide and the second leading cause of cancer-related deaths globally [1, 2]. In China, colorectal cancer ranks second and fourth, respectively, in terms of incidence and mortality among all malignant tumors [3].

Since the disease often presents no obvious symptoms in its early stages and can be easily mistaken for other conditions, the lack of effective screening methods means that most patients are diagnosed at advanced stages, missing the optimal window for treatment and facing a serious threat to their health. Therefore, early screening and early diagnosis can effectively prevent the development of colorectal cancer. Although tools such as colonoscopes can provide clear images of polyps, there are still many cases of misdiagnosis and missed diagnoses. Clinical diagnosis heavily relies on physicians' experience, and variations in experience introduce uncertainty. Polyps differ significantly in size and shape, and their low contrast with surrounding tissue makes boundaries difficult to discern. Based on these issues, it is necessary to develop an automated tool for segmenting colorectal polyps, which can improve diagnostic ability while reducing medical costs.

Computer-aided automatic segmentation went through two stages: the traditional machine learning stage and the deep learning stage. In the traditional machine learning stage, medical image segmentation mainly relied on methods such as thresholding, region growing, edge detection, and morphological operations, using manually designed features to extract and segment target regions. These methods classify pixels or segment regions by analyzing low-level features such as intensity, gradient and texture, combined with manually designed decision rules. However, due to their limited feature representation and sensitivity to noise and imaging conditions, their segmentation performance is often constrained in complex medical image scenarios. In the deep learning stage, convolutional neural networks (CNNs) can automatically learn abstract feature representations from large annotated images and fit the true probability distribution of pixels, enabling end-to-end segmentation of target regions. For example, in the classic UNet [4] model, the encoder progressively extracts semantic features from the image, while the decoder gradually restores them to the original resolution with the help of skip connections, achieving accurate segmentation. Many U-shaped network architectures have since emerged, such as ResUNet++ [5], densely connected UNet++ [6], and PraNet [7] with reverse attention. Most of these focus on improving learning paradigms, including feature fusion and attention mechanisms [8-10].

With the introduction of the Transformer [11] architecture, self-attention–based models have gradually been applied to computer vision, with typical examples including Vision Transformer (ViT) [12] and Swin Transformer [13].

These methods use self-attention to model long-range dependencies and can effectively capture correlations between different regions. However, their ability to extract local details is limited, which restricts their full potential in the visual domain. To address this issue, attempts have been made to combine Transformers with convolutional neural networks. For example, networks like TransUNet [14] and Transfuse [15], combine CNNs and Transformers, enhancing the modeling of long-range dependencies while preserving the ability to capture detailed information.

In recent years, Mamba [16] has gradually garnered attention from the research community. Known for its linear time complexity and long-range dependency modeling capabilities comparable to those of Transformers, Mamba has

also been extensively explored in the field of medical image segmentation. VM-UNet [17] uses Visual State Space(VSS) blocks to replace convolutional layers, which effectively models long-range dependencies and achieves good segmentation performance. Polyp-Mamba [18] employs Mamba and convolutional neural networks for feature extraction and fusion respectively. It not only captures local details effectively but also models the correlation between different regions.

The following sections of this paper focus on existing contributions to colorectal polyp segmentation, reviewing them in terms of network architectures, learning paradigms, and supervision methods. Next, commonly used datasets and evaluation metrics for colorectal polyp segmentation are introduced. Finally, the main challenges of deep learning–based polyp segmentation and potential directions for future development are discussed.

2. Methodology

2.1. Network Architectures

2.1.1. Encoder-Decoder Architectures

In the field of deep learning–based image segmentation, encoder–decoder architectures have become the mainstream and highly influential due to their simple and effective design. The encoder extracts semantic information from the image in a hierarchical manner, while the decoder progressively reconstructs the segmentation mask in a cascaded fashion.

The Fully Convolutional Network (FCN) [19] pioneered end-to-end pixel-wise prediction, but it suffers from detail loss, resulting in suboptimal segmentation performance.

The UNet architecture is further optimized and improved on the basis of FCN. To compensate for the loss of spatial detail information caused by downsampling, UNet concatenates the features output by the encoder to the corresponding decoder at the same layer, which effectively alleviates the problem of missing local details in segmentation. Fig. 1 shows the structure of UNet.

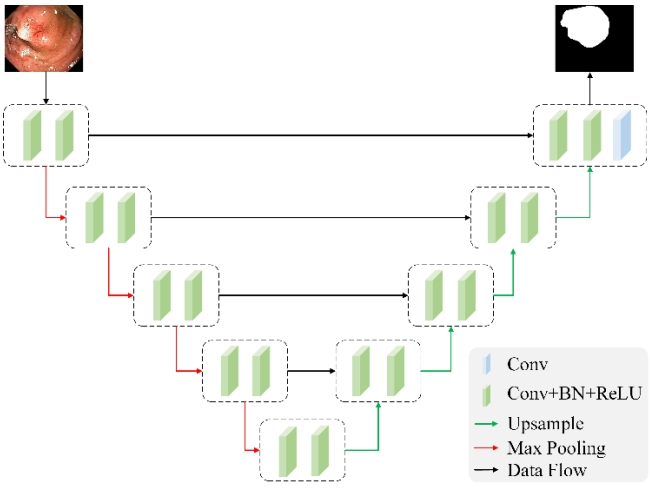

Fig. 1 UNet network architecture.

Attention UNet [20] introduces attention mechanisms at the skip connections, allowing the network to focus more on important regions in the image and improve segmentation accuracy. UNet++ further enhances feature fusion by designing more complex nested skip connections.

SegNet [21] uses Max Pooling indices recorded during the encoder stage for nonlinear upsampling in the decoder, effectively preserving spatial structure while reducing the number of parameters.

ResUNet++ extends the traditional UNet by adding residual connections, attention mechanisms, and an Atrous Spatial Pyramid Pooling (ASPP) module to better capture complex boundaries and fine structures. Residual units improve feature extraction, attention gates strengthen key feature transfer, and ASPP provides multi-scale context, enhancing segmentation performance. [22] adopts the DenseNet [23] design and incorporates attention mechanisms to model long-range dependencies across spatial and channel dimensions, enhancing the fusion of spatial and semantic information, reducing redundant features, and bridging the semantic gap between the encoder and decoder to improve segmentation performance.

2.1.2. Transformer-based Architectures

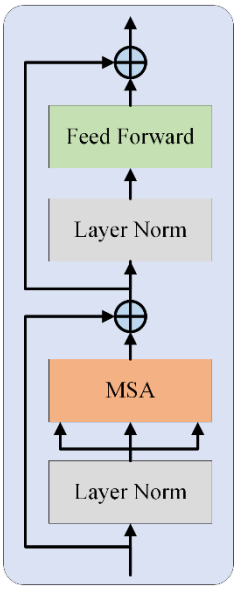

Fig. 2 Transformer Encoder Layer

Although CNNs have achieved great success in medical image segmentation, their local receptive fields make it difficult to model long-range dependencies, which limits segmentation performance in complex scenarios. The introduction of Transformer [11] has significantly advanced image modeling, as its self-attention mechanism can capture global semantic relationships across different regions of an image. Originally developed for natural language processing, the Vision Transformer (ViT) [12] applies a Transformer architecture to image tasks, demonstrating its potential in the visual domain and opening new directions for medical image segmentation. Fig. 2 shows a single layer of the Transformer encoder, which can be used for visual feature extraction.

Subsequently, TransUNet [14] enhances medical image segmentation by combining the global dependency modeling capability of Transformers with the local detail extraction capability of CNNs, but its high computational complexity and large number of parameters make it heavily dependent on data volume and computing resources. TransFuse [15] captures local details and global semantics using CNN and Transformer encoders, respectively, and integrates semantic and spatial information through a fusion mechanism, thereby improving segmentation accuracy.

2.1.3. Mamba-based Architectures

In recent years, Mamba-based network have gradually become popular in the field of medical image segmentation. [24] applies the Mamba model to polyp segmentation,

efficiently capturing long-range dependencies while keeping linear computational complexity and enhancing cross-scale semantic relationships. Polyp-Mamba [18] combines a Mamba branch with a CNN branch for feature extraction to address the challenges of high similarity between polyps and surrounding tissue and unclear boundaries. [24, 25] replace both the encoder and decoder of UNet with VSS blocks to improve segmentation performance, while SAM-Mamba [26] introduces a Mamba prior module into SAM to capture global context, achieving superior zero-shot segmentation capability.

2.1.4. Dual-branch Network

Dual-branch networks can be classified based on their structure into dual-branch encoders, dual-branch decoders, and hybrid dual-branch designs. These architectures have all been widely used in colorectal polyp segmentation tasks. Dual-branch encoder architectures typically design two heterogeneous or homogeneous encoder branches, which are responsible for extracting global and local features, or deep semantic and shallow detail features, respectively. For example, BiSeNet [27], as a type of dual-branch encoder, uses two parallel CNN branches to extract contextual information and spatial information, respectively.

In addition, CNNs are often combined with Transformers or Mamba. In such cases, the CNN branch is typically used to capture local spatial features of polyps, while the Transformer or Mamba models long-range dependencies across different regions through self-attention, achieving a complementary representation of local details and global context. In colorectal polyp segmentation, this type of architecture can effectively address issues such as blurred polyp boundaries and low contrast with surrounding tissue. For example, TransFuse [15] effectively uses CNN and Transformer dual branches to extract rich detailed and abstract semantic information.

HMT-UNet [28] explores the effectiveness of a hybrid framework combining SSM and Transformers for polyp segmentation. In addition, dual-decoder architectures have also been applied to polyp segmentation tasks. For instance, DG-Net [29] employs two decoder branches: one to enhance the semantic distinction between polyps and the background, and the other to explicitly model the target boundary information, thereby improving boundary localization accuracy by jointly modeling regional semantic and boundary structure information within the dual-decoder framework.

2.2. Learning Paradigm

2.2.1. Multi-scale Fusion

Multi-scale feature fusion is an important research direction in semantic segmentation, aiming to model image semantic information at different spatial scales to address the segmentation challenges caused by variations in object size. To this end, many methods adopt a pyramid-like feature hierarchy, extracting and fusing features at different resolutions, where high-resolution layers preserve rich spatial details and low-resolution layers provide stronger semantic context, thereby forming multi-scale representations that balance local details and global semantics.

The Feature Pyramid Network (FPN) [30] is one of the representative methods. It constructs a feature fusion pathway combining bottom-up and top-down processing and employs lateral connections to merge features across different levels, enabling high-level semantic information to be effectively propagated to high-resolution feature maps. This results in semantically consistent feature representations across all scales, significantly enhancing the model's ability to detect and segment objects of varying sizes.

DeepLabV3 [31] uses dilated convolutions to enlarge the receptive field and introduces ASPP module, which performs multi-scale feature sampling through parallel convolutions with different dilation rates. This allows effective capture of context information at multiple scales while maintaining computational efficiency and improving segmentation performance. PSPNet [32] proposes the Pyramid Pooling Module (PPM), which applies pooling windows of different scales on feature maps to aggregate global context information across multiple scales. PPM extracts regional features from various spatial scales and fuses them with the original features via upsampling, thereby enhancing the model's ability to represent global context and improving segmentation performance in complex scenarios.

2.2.2. Attention Mechanism

In colorectal polyp segmentation tasks, polyp regions often exhibit varying scales and shapes, as well as blurred boundaries, and the images are frequently affected by challenging conditions such as abnormal exposure. Therefore, relying solely on the local receptive fields of convolutional networks is often insufficient to capture critical discriminative information. To address these issues, many studies have incorporated attention mechanisms into segmentation models, which adaptively weight features to make the network focus more on information relevant to the target regions, thereby improving segmentation accuracy.

Early attention mechanisms typically model features along the spatial or channel dimensions [33, 34]. Spatial attention emphasizes key locations in the image, enabling the model to focus more on potential target regions, while channel attention models the importance of different feature channels, assigning higher weights to channels with stronger semantic representation to enhance feature expressiveness. For example, [5] employs a squeeze-and-excitation mechanism in the encoding stage to strengthen the representation of important feature channels, and [2] designs a parallel attention mechanism that uses reverse attention to guide feature selection, effectively capturing polyp boundary information.

As research has progressed, attention mechanisms have evolved from reweighting local features to structures that can model global dependencies, such as self-attention [11] and more recent linear-complexity sequence modeling methods like Mamba [16]. These methods can capture long-range relationships between pixels or regions across the entire image, helping overcome the limited local receptive field of convolutional networks and significantly improving the model's ability to represent complex structures and global semantic information.

In recent years, some studies have introduced attention mechanisms in the frequency domain to better utilize the global structure of images. By modeling both low- and high-frequency components, networks can more accurately capture the contours and texture features of polyps, improving the segmentation of small or blurry-boundary polyps. For example, [18] uses a multi-frequency perception module to capture both local and global features, while a spatial attention module guides the model to focus on key regions, enabling more accurate extraction of polyp boundaries and shapes and enhancing segmentation performance.

2.2.3. Diffusion Model

Diffusion models, as a class of generative networks that

have attracted significant attention in recent years, demonstrate strong generative capabilities by deeply modeling data distributions. Their core idea is to model data progressively through two stages: forward diffusion and reverse sampling. In the forward diffusion process, noise is gradually added to the image, smoothing the data distribution; in the reverse sampling process, the model progressively restores the image by predicting the noise, achieving high-quality generation. Introducing diffusion models into image segmentation tasks can enhance the model's understanding of the intrinsic structure and semantic relationships in the data, leading to more accurate and detailed segmentation results. SegDiff [35] was the first to apply diffusion models to image segmentation, using the input image as a guidance condition for semantic segmentation, allowing the model to learn the input image's data distribution and perform segmentation during the stepwise denoising process. Wolleb et al. [36] further improved segmentation performance by evaluating the segmentation uncertainty of diffusion models.

2.3. Levels of Supervision

In colorectal polyp image segmentation, model training approaches can generally be divided into three categories. In supervised learning, models are trained on polyp segmentation datasets with pixel-level annotations to achieve high segmentation performance. In semi-supervised learning, a small amount of labeled data is combined with a large set of unlabeled images to improve segmentation performance. In self-supervised or unsupervised learning, models are pre-trained using self-supervised methods and then transferred to segmentation tasks to reduce reliance on annotated data.

3. Polyp Segmentation Datasets

Deep learning models essentially learn the underlying probability distribution from large amounts of existing data, so their training typically relies on high-quality datasets. With the development of polyp segmentation research, a variety of datasets have become available [37]. Table 1 lists some commonly used polyp-related datasets and their key characteristics.

Table 1. Public Polyp Segmentation Datasets

Datasets	Number	Resolution	#Obj.
CVC-ClinicDB	612	384×288	1~3
CVC-ColonDB	300	574×500	1
EndoScene	912	384×288 to 574×500	1~3
Kvasir-SEG	1,000	332×487 to 1920×1072	1~3
Kvasir-sessile	196	401×415 to 1348×1070	1~3
BKAI-IGH	1,200	1280×959	1~18

4. Evaluation Metrics

To evaluate the segmentation performance of models on colorectal polyps, this study lists several commonly used metrics.

The Dice coefficient is used to evaluate the similarity between the predicted mask and the ground truth.

$$Dice=\frac{2TP}{2TP+FP+FN} \tag{1}$$

The Intersection over Union(IoU) is used to measure the overlap between the predicted mask and the ground truth mask. It is calculated as follows:

$$IoU=\frac{TP}{TP+FP+FN} \tag{2}$$

Recall represents the proportion of true positive samples that are correctly identified by the model.

$$Recall=\frac{TP}{TP+FN} \tag{3}$$

Precision measures the proportion of predicted positive samples that are actually correct.

$$Precision=\frac{TP}{TP+FP} \tag{4}$$

Accuracy represents the proportion of all pixels that are correctly classified by the model.

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

5. Challenges and Future Directions

5.1. Unsupervised Learning

Currently, most polyp segmentation models are trained using supervised learning. However, due to the scarcity of high-quality annotated datasets, supervised methods are prone to prediction bias when data distributions change. Moreover, models trained on small datasets often suffer from overfitting, limiting their ability to generalize to real-world scenarios. In contrast, unsupervised or weakly supervised learning approaches can leverage unlabeled or weakly labeled polyp images, using large-scale data to mitigate distribution shifts caused by limited annotations, thereby enhancing model robustness and generalization.

5.2. Lightweight Models and Applications

Although existing deep networks achieve high segmentation accuracy, increasing model parameters to boost performance results in high computational costs, making them unsuitable for mobile or embedded devices and limiting their clinical applicability. Future research should focus on designing lightweight architectures that balance performance and efficiency. Additionally, optimizing inference speed, memory usage, and model stability, as well as exploring end-to-end integration with clinical systems, will facilitate the deployment and practical use of polyp segmentation technology in real-world medical settings.

5.3. Interactive Models

In recent years, general segmentation models such as the Segment Anything Model have shown great potential in human-guided segmentation. By combining limited interactive annotations from clinicians with powerful pre-trained models, segmentation efficiency and accuracy can be significantly improved. This approach not only reduces the burden of manual annotation but also ensures the reliability of clinical decision-making, representing an important direction for future research in polyp segmentation.

6. Conclusion

This paper provides a comprehensive review of existing research on colorectal polyp segmentation. To clearly illustrate recent innovations, we discuss polyp segmentation

methods in terms of network architectures, novel paradigms, and training strategies. We then introduce commonly used datasets in this field and briefly describe standard metrics for evaluating segmentation performance. Finally, we explore potential directions for future automated polyp segmentation methods. By systematically reviewing existing approaches and analyzing the key challenges in current polyp segmentation tasks, this work aims to provide researchers with insights into the forefront of innovation in the field.

Acknowledgements

Funding: This work was supported by National Natural Science Foundation of China under Grant No.62276092, 62303167, Science and Technology Research of Henan Province No. 252102210042.

References

[1] J. Silva, A. Histace et al., "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," vol. 9, no. 2, pp. 283-293, 2014.

[2] X. Kang, Z. Ma et al., "Multi-scale information sharing and selection network with boundary attention for polyp segmentation," vol. 139, p. 109467, 2025.

[3] Y. Liu, C. Zhu et al., "Temporal trends in disability adjusted life year and mortality for colorectal cancer attributable to a high red meat diet in China from 1990 to 2021: an analysis of the global burden of disease study 2021," BMC Gastroenterology, vol. 24, no. 1, p. 476, 2024/12/27 2024, https://doi.org/10.1186/s12876-024-03563-7.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234-241: Springer.

[5] D. Jha, P. H. Smedsrud et al., "Resunet++: An advanced architecture for medical image segmentation," in 2019 IEEE international symposium on multimedia (ISM), 2019, pp. 225-2255: IEEE.

[6] Z. Zhou, M. M. Rahman Siddiquee et al., "Unet++: A nested u-net architecture for medical image segmentation," in International workshop on deep learning in medical image analysis, 2018, pp. 3-11: Springer.

[7] D.-P. Fan, G.-P. Ji et al., "Pranet: Parallel reverse attention network for polyp segmentation," in International conference on medical image computing and computer-assisted intervention, 2020, pp. 263-273: Springer.

[8] R. Tang, H. Zhao et al., "A frequency attention-embedded network for polyp segmentation," vol. 15, no. 1, p. 4961, 2025.

[9] T. Zhou, Y. Zhou et al., "Cross-level feature aggregation network for polyp segmentation," vol. 140, p. 109555, 2023.

[10] R. Zhang, P. Lai et al., "Lesion-aware dynamic kernel for polyp segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 99-109: Springer.

[11] A. Vaswani, N. Shazeer et al., "Attention is all you need," vol. 30, 2017.

[12] A. Dosovitskiy, L. Beyer et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

[13] Z. Liu, Y. Lin et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012-10022.

[14] J. Chen, Y. Lu et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[15] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in International conference on medical image computing and computer-assisted intervention, 2021, pp. 14-24: Springer.

[16] A. Gu and T. J. a. p. a. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023.

[17] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," 2024, https://doi.org/10.48550/arXiv.2402.02491.

[18] X. Zhu, W. Wang et al., "Polyp-mamba: A hybrid multi-frequency perception gated selection network for polyp segmentation," vol. 115, p. 102759, 2025.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440, https://doi.org/10.1109/CVPR.2015.7298965.

[20] O. Oktay, J. Schlemper et al., "Attention u-net: Learning where to look for the pancreas," 2018, https://doi.org/10.48550/arXiv.1804.03999.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," (in English), Ieee Transactions on Pattern Analysis and Machine Intelligence, Article vol. 39, no. 12, pp. 2481-2495, Dec 2017, https://doi.org/10.1109/tpami.2016.2644615.

[22] T. Hussain, H. Shouno et al., "DCSSGA-UNet: Biomedical image segmentation with DenseNet channel spatial and semantic guidance attention," vol. 314, p. 113233, 2025.

[23] G. Huang, Z. Liu et al., "Densely Connected Convolutional Networks," arXiv [cs.CV], 2018 2018.

[24] Z. Xu, F. Tang et al., "Polyp-mamba: Polyp segmentation with visual mamba," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024, pp. 510-521: Springer.

[25] H. Yan, Q. Hong et al., "SCM-UNet: Spatial-channel Mamba UNet for medical image segmentation," Digital Signal Processing, vol. 168, p. 105550, 2026/01/01/ 2026, https://doi.org/https://doi.org/10.1016/j.dsp.2025.105550.

[26] T. K. Dutta, S. Majhi et al., "SAM-Mamba: Mamba Guided SAM Architecture for Generalized Zero-Shot Polyp Segmentation," arXiv [cs.CV], 2024 2024.

[27] C. Yu, J. Wang et al., "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 325-341.

[28] M. Zhang, Z. Chen et al., "HMT-UNet: A hybird Mamba-Transformer Vision UNet for Medical Image Segmentation," arXiv [eess.IV], 2024 2024.

[29] D. He, Y. Li et al., "Dual-guided network for endoscopic image segmentation with region and boundary cues," vol. 91, p. 106059, 2024.

[30] T.-Y. Lin, P. Dollár et al., "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[31] L.-C. Chen, G. Papandreou et al., "Rethinking atrous convolution for semantic image segmentation," 2017.

[32] H. Zhao, J. Shi et al., "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881-2890, https://doi.org/10.1109/CVPR.2017.660.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[34] S. Woo, J. Park et al., "CBAM: Convolutional Block Attention Module," arXiv [cs.CV], 2018 2018.

[35] T. Amit, T. Shaharbany et al., "Segdiff: Image segmentation with diffusion probabilistic models," 2021.

[36] J. Wolleb, R. Sandkühler et al., "Diffusion models for implicit image segmentation ensembles," in International Conference on Medical Imaging with Deep Learning, 2022, pp. 1336-1348: PMLR.

[37] Z. Wu, F. Lv et al., "Colorectal polyp segmentation in the deep learning era: A comprehensive survey," 2024, https://doi.org/10.48550/arXiv.2401.11734.